Phylogenetics in a Modern-Day Pandemic

Nicholas Chan

This paper covers an analysis and comparison of various tree placement programs in phylogenetics and their usage in contact tracing during the 2019-2021 COVID-19 pandemic. This paper will begin with laying out the background of phylogenetics and the idea of maximum parsimony in order to describe the issue imposed by COVID-19, a rapidly mutating and highly transmissible virus at the center of a global pandemic. This paper will then discuss how the highest performing programs have been used for contact tracing. The intended audience of this paper are those interested in the applications of contact tracing or those who have practiced tracing lineages of organisms with bioinformatics tools. This paper is written according to the IEEE style guide.

1 ABSTRACT

DATABASE keeping and management of genomics data is an interdisciplinary practice that originated in the early 80's as a marriage between phylogenetics and simple computer science algorithms for string parsing. In just the last year, active genome databases around the globe have seen an unprecedented growth in size as the world entered the COVID-19 pandemic. With such an uptake of sequencing data equating to over half of the total number sequences submitted since 1982 occurring in just 2020, growth seems unlikely to falter as sequencing techniques and bioinformatics software continuously improve to more rapidly and precisely produce data. The purpose of this paper is to inform the reader of how phylogenetics analyses are performed today and how implementations of such algorithms in the form of programs work to track emerging evolutionary pathways of COVID-19 during this pandemic. This paper will also provide a comparative analysis of several programs used in phylogetics analysis and how well they serve to aid the global contact tracing effort.

2 INTRODUCTION

Phylogenetics is the study of the evolutionary histories and relationships between different classes of organisms, otherwise known as species. The origins of phylogenetics date back to the Genomics Era of the mid-twentieth century which had been characterized by discovery of the Watson-Crick base pairing scheme of DNA. The revolution that was the discovery of the Watson-Crick base pairing scheme of DNA uprooted many false hypotheses on heredity and evolution to established the foundation for modern genomics. For the field of phylogenetics, this meant that evolution and lineages could be tracked and predicted by data beyond the physical appearances or habitats of organisms. Phylogenetic studies could now be grounded in the existence of discrete units which came in the form of nucleic acids. Sequences of nucleic acids provided identifying information as well as the ability for researchers to track an organism's genetic signature with molecular precision simply based on the characteristics of its genome sequence. The use of sequencing data as a means of identification became known as DNA testing. However, with all the guiding principles necessary for identifying organism evolution now layed out, the only issue remaining was the lack of readable sequences of DNA. Fast forward to today in the Postgenomic Era of the early twenty-first century, whole genome sequencing data has become more available than it has ever been, with its domain now closer to computers than traditional wet labs. By 2005, the shift in the domain of genetic data had enabled the rate at which biological data was dispersed and produced to accelerate in a manner only comparable to Moore's law. This was facilitated in large by online databases owned by institutions in both academia and industry that housed the majority of high throughput sequencing data with sequencing technology also having advanced at such a rate where timely and affordable whole genome sequencing was possible. However, even with such technological change over half a century in the field of phylogenetics, its founding principles remained. One such principle was maximum Parsimony.

3 MAXIMUM PARSIMONY

Maximum parsimony is the main concept used in phylogenetics for deciding the most plausible lineage of an organism's evolution. Maximum parsimony, in summary, states that with many outcomes, the most likely outcome would possess the fewest changes. In terms of phylogenetics, the most optimal lineage wioould be the sequence of evolutions where the least number of mutations occurred. In other words, the sequence of events where the total number of nucleotide substitutions was minimized. The total count of substitutions is called the parsimony score and acts as a metric to a tree's overall complexity. To better illustrate this idea, refer to Figure 1, which compares two phylogenetic trees with different parsimony scores. From this figure you should be able to notice that a lower score is linked to a tree being more parsimonious and thus more simple and likely.

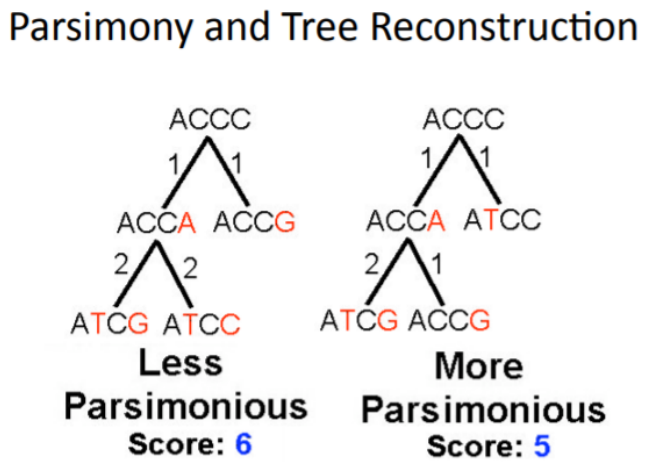

Fig. 1. Illustration of a tree graph abstract data type using parsimony score as the edge weight metric.

A higher parsimony score exhibits the opposite relation. You will also notice that the nodes of this tree graph are occupied by sequences of nucleotides. In reality, these sequences would be around the magnitude of a hundred thousand to a million times larger than the mere four to five nucleotide long sequences depicted. Despite the difference between the true and figurative sizes of sequences considered for maximum parsimony, the method of how parsimony score is calculated remains the same for both. One of the earliest and most famous algorithms based on the idea of maximum parsimony was designed by Walter M. Fitch in 1971.

4 Algorithms for Calculating Parsimony Statistics

Two algorithms based on the idea of maximum parsimony that have risen to the top of modern computational phylogenetics are the Fitch and Sankoff algorithms. These two algorithms have been used to great extent in many tree sample placement software tools in use today for the analysis of phylogenetic trees and lineages of organisms. Both algorithms are essential in computational phylogenetics due to their robustness and reliability granted by the heuristics each of them employ to account for real world noise that commonly occurs in sequencing data.

4.1 The Fitch Algorithm

One of the earliest and most famous algorithms based on the idea of maximum parsimony was designed by Walter M. Fitch in 1971[1]. The Fitch algorithm is a heuristic used for finding an optimal tree given the same type of multiple sequencing data that was described to have occupied the nodes of the tree graph from Figure 1. The main feature of the Fitch algorithm is its robustness for determining genetic distances among multiple sequencing data, which are often distantly related. A possible cause for this phenomenon is simply because of the times that sequences were produced or other wet lab errors that may have confounded the sequencing data. To complete such a task with such varied data occupying the nodes of the graph, a substitution matrix, otherwise known as a distance matrix in computer science, is created via least squares method to better determine the actual genetic distance between the sequences held at each node. An example of this matrix can be seen in Figure 2, which illustrates how certain substitutions of nucleotides are given higher or lower parsimony scores. In summary, the substitution matrices in the Fitch algorithm are used to calculate parsimony scores of trees in a way that accounts for inaccuracies observed through lab findings pertaining to how the multiple sequencing data was sequenced. As a result, the Fitch algorithm is a great choice for determining the parsimony and placement of an optimal tree.

	B	C	D	E
A	.31	1.01	.75	1.03
B	-	1.00	.69	.90
C	-	-	.61	.42
D	-	-	-	.37

Fig. 2. Example of the values the Fitch algorithm substitution matrix could take in order to account for data uncertainty[9].

4.2 Sankoff Algorithm

The second out of the two algorithms that are to be discussed is the Sankoff algorithm, which was developed by David Sankoff in 1990[2]. The Sankoff algorithm is a heuristic similar in purpose as the Fitch algorithm in the way it seeks tree graphs with the smallest weights associated with them. However, the Sankoff algorithm creates and utilizes the scoring matrix centered around penalizing the scores associated with nucleotide substitutions on gaps and mismatches between the sequences held at nodes on a graph. Finding a tree with the minimal score as defined by the Sankoff algorithm would also mean finding the tree where the least amount of substitutions on gaps or mismatches occurred. Although the Sankoff algorithm relies on a different heuristic for cancelling noise associated with multiple sequencing data compared to what the Fitch algorithm uses, the Sankoff algorithm otherwise produces practically identical results as the Fitch algorithm for determining the placement of an optimal tree.

5 The Fitch and Sankoff Algorithm in Software

With the framework for determining the placement of optimal trees from sequencing data mostly laid out by the Fitch and Sankoff algorithms, it is not a surprise that they have seen frequent use in implementation of tree sample placement today. Such programs that the bioinformatics community considers benchmarks in tree sample placement are PAGAN2 (developed by the University of Helsinki)[3], IQ-TREE2 (developed by the University of Vienna)[4], TreeBeST2 (developed by the Chris Ponting group)[5], and RAxML epa (developed by the Exelixis Lab)[6]. However, the most recent and promising development among these is UShER, which was developed by Professor Corbett-Detig's lab at UCSC just last summer[8]. To briefly describe each program mentioned and a design feature unique to them, the main use of the PAGAN2 program is for inferring ancestral sequences from a tree graph built from a reference sequence. IQ-TREE2 is a program used to capture aspects of genomics sequence evolution through a stochastic algorithm adapted from the Fitch and Sankoff algorithms for inferring phylogenetic trees by maximum likelihood. TreeBeST2 is a program that builds gene trees from known species phylogenies in the initial phase of its algorithm and then calculates a probable gene tree in its closing phase where neighbor-joining of nodes takes place. RAxML is a program that runs on an evolutionary placement algorithm which calculates edge weights based on the likelihood statistics model for maximum parsimony. UShER uses the traditional principles of Maximum Parsimony as well as a combination of the scoring matrix schemes used by the Fitch and Sankoff algorithms according to its documentation.

6 Choosing the Optimal Program

With so many different implementations of programs that often yield near identical for tracking the optimal placement of tree samples, choosing a single program to use can become a cumbersome task. To remedy this issue, I decided to run a comparative analysis of these programs and report my findings on which programs performed the best.

6.1 Comparative Analysis of Program Performances

The test that I conducted was a simple performance test timing how long each of the Tree sample placement programs in my comparative analysis took to place 1000 sequences onto a tree. The programs that I used in my comparative analysis were PAGAN2 , IQ-TREE2 , TreeBeST2 , and RAxML epa. For programs where I was able to enable time recording I recorded the time reported and for programs where timing a session was not an option, I timed the run manually with a stopwatch. These tests were performed on a laptop with an i5 processor.The sequencing data used in my comparative analysis of Tree Sample Placement Softwares was a sub-sample of 1000 SARS-CoV2 genome sequences sourced from the GISAID database, which holds datasets on of

COVID-19 variants sequenced in labs around the world and has partnered with the Nexstrain database. This experiment was inspired by the paper Stability of SARS-CoV-2 Phylogenies by Professor Corbett-Detig's lab and Yatish Turakhia which tested the performances of these same programs in a similar fashion.

Method	Time to Place 1000 Sequences
PAGAN2	24+ Hours
IQ-TREE2	24+ Hours
TreeBeST	24+ Hours
RAxML epa	24+ Hours
UShER	43.2 Seconds

Fig. 3. Results from comparison of run times for placing 1000 sequences onto a phylogenetic tree

6.2 UShER's Optimization for Tree Sample Placement

To understand some of the causes for the disparity in performance between UShER and the rest of the programs tested, I will first discuss the underlying algorithms used in tree sample placement. To reiterate, the driving idea behind coming up with the most likely lineage or the most likely path that SARS-CoV2 took in evolution is maximum parsimony. Maximum parsimony is a principle stating that the simplest turn of events is the most likely. The concept of maximum parsimony is interpreted in the context of computer algorithms and phylogeny as a metric for measuring minimal levels of character changes observed in a tree. Trees with the lowest possible parsimony are the simplest and thus the most likely to not occur by chance. As a result of this relationship between individual sequences and the parsimony of a lineage, the core data structures used in these programs are acyclic graphs, or trees. The nodes of these trees are represented by sequencing data while the edge weights are represented by the differences found between the sequences of nodes. The main algorithms used for this are the Sankoff and Fitch algorithms. With the Sankoff algorithm, you start off with the root vertex and a set of parsimony scores for each of the 4 possible nucleotides, A, C, T, and G. A substitution matrix is also used in finding the most likely path of mutation. Substitution matrices are usually determined by lab findings. A common finding is that since the gene you're looking at is a primer binding site, mutations from C to T are more likely and would thus have a lower parsimony score. Going down from the root node, children nodes are assigned an optimal character based on the nucleotide with the lowest parsimony score. Up to this point, these have been the general design features used by the tree placement programs, excluding UShER. Where UShER differs from most previous tree placement programs is in its design choice to utilize positional data of substitutions in multiple sequencing data rather than whole multiple sequencing data itself which most similar programs have opted to use. The positional data used by UShER is a new take on how the Fitch and Sankoff algorithms were utilized as the sequencing data that traditionally occupied the nodes of the tree graphs in those algorithms was now replaced by a more compressed format of data known as VCF or variant call format. Because of this design choice, UShER can perform much faster than most tree placement programs in existence which presumably still use the traditional design choice of using whole multiple sequencing data. The consequence of this is an average sample placement time of 1/10th of a second that accurately performs 97% of the time according to UShER's documentation. Although these programs varied greatly in terms of runtime, they all followed the same core principles of parsimony. UShER however was optimized for speed. UShER's implementation utilized metadata of sequences rather than brute forcing the Fitch-Sankoff algorithm through a data set of 1000 Sars CoV2 variant sequences, each around 30000 characters long. As a result, UShER is the most optimal program for phylogenetic tree construction/addition when using sequencing data until. Although all benchmark programs that were tested against UsHER used either the Fitch or Sankoff algorithms, UsHER's used of both of them as well as a preprocess-

ing phase when reading input allowed it to surpass the others in speed. The preprocessing phase in UsHER converts normal sequencing data files to variant call format (VCF) which is used to convey metadata. The conversion of sequencing data to a more compressed format significantly cuts run time.

7 Applications of Tree Placement Software in the COVID-19 Pandemic

As a result of the utility offered by tree sample placement software in lineage building and reconstructing likely evolutionary pathways, such programs have found use in contact tracing during the COVID-19 pandemic. The global contact tracing initiative has exponentially increased the rate that we have sequenced viral genomes, with an uptake of sequencing data equating to over half of the total number of sequences submitted since 1982 occurring in just 2020. The issue that my final project addresses is how sample placement programs must be tailored in order to provide useful inferences from virus lineages such as its mutation trajectory and path of spread and which published programs work the best. Nearly all sample placement programs in use today were developed prior to the pandemic and haven't been optimized for the purposes required by contact tracing. As a result these programs sorely underperformed when tackling the volume of sequencing data needed for contact tracing today. See Figure 3 for the consensus time these programs took to run through a simulated lineage tracing experiment with 1000 sequences. However, with the goals of contact tracing and its need for rapidly processed data on viral lineages well known to scientists studying the pandemic, programs such as UShER were developed with those goals as its guiding principles for its design.

7.1 UShER's Usage in Genomic Contact Tracing

In just the past year, UShER has seen two ports as an online tool for identifying genetic similarities among newly sequenced viral genomes. With its unmatched ability to quickly determine optimal tree placements and lineages, UShER has left the genomic contact tracing powerhouses known as Next Strain Database and the UCSC Genome Browser no better alternative for tree sample placement applications in tracking emerging SARS-CoV2 lineages. On each of these sites, UShER can be accessed and run in real time, providing great utility for epidemiologists and health care specialists in the spread and evolution of SARS-CoV2. Training modules for the use of UShER on these platforms have also been made available by the CDC as a part of their Official COVID-19 Genomic Epidemiology Toolkit for understanding genomic contact tracing. Unlike conventional contact tracing methods where the spread of SARS-CoV2 has been tracked by human resources and reported through communicative means such as smartphone alerts, genomic contact tracing tracks the evolutionary distances between viral genomes sequenced from varying communities. In genomic contact tracing, the main indicator of viral spread is a statistically significant level of genetic similarity observed between viral genomes sequences in neighboring communities. Non statistically significant or less statistically significant levels of genetic similarity between viral genomes sequences among neighboring communities indicate a more distant relationship and a lower chance that frequent transmission of the virus had recently occurred.

7.2 UShER's Port as a Webtool on the Nextstrain Database

Nextstrain is considered as one of the progenitors of what modern viral genome databases were to become, a highly accessible database and web tool with a dynamic library of sequencing data. However, in its infancy, Nextstrain was an open-source collaboration between labs around the world with the goal of establishing a network capable of tracking pathogens and outbreaks. Similar databases and web tools have existed for a relatively long time prior to Nextstrain's arrival, but Nextstrain main claim to fame was this new concept of real time tracking of heavily se-

quenced pathogens[7]. Nextstrain features coupled functionality with the GISAID database from where most of its reference sequences are sourced from as well the Virus Pathogen Resource Database (ViPR) from where many of Nextstrain's curated viral and archaeal genome sequences come from. Nextstrain's proof of the viability of this concept lay in its humble beginnings as a visualizer for the spread and mutation frequency of Influenza, West African Ebola, Dengue, Measles, Mumps, Tuberculosis, West Nile virus and Zika. Soon after the dawn of the SARS-CoV2 pandemic in December 2019, Nextstrain became a centerpiece of both sequencing data queries and submissions. Queries run through Nextstrain take new genome sequences of an organism as input and output a phylogenetic tree built with a reference genome sequence as its root. With a figure as simple as this phylogenetic tree, a map of the spread of a virus can be drawn out on a global map, with the option to highlight variants in different colors. This feature has been coined as joint temporal and spatial visualization, because of Nextstrains ability to illustrate evolution through space(the world) and time. However, Nextstrain's role as a sequencing database has been reduced due to its partnership with GISAID, a more widely known public repository for sequencing data that has seen its own unprecedented uptake in SARS-CoV2 related datasets. UShER's port on the Nextstrain site functions to construct plausible lineages while taking advantage of Nextstrain's capabilities in joint temporal and spatial visualization. The webtool born out of the combination of these two interfaces enables users to add additional samples onto trees annotated with plausible lineages to simulate where newly sequenced viral samples would likely fall in relation to a previously placed lineages. An example of such a visualization can be seen in Figure 4, which depicts a joint temporal and spatial visualization of the 20A, 20C, 20B, 20E, 20G and 20H SARS-CoV2 variant lineages. In Figure 4, UShER's utility in Ultra Fast sample placement is exhibited by the addition of sequencing samples, highlighted in red,onto the 20C SARS-CoV2 lineage of the known SARS-CoV2 phylogeny.

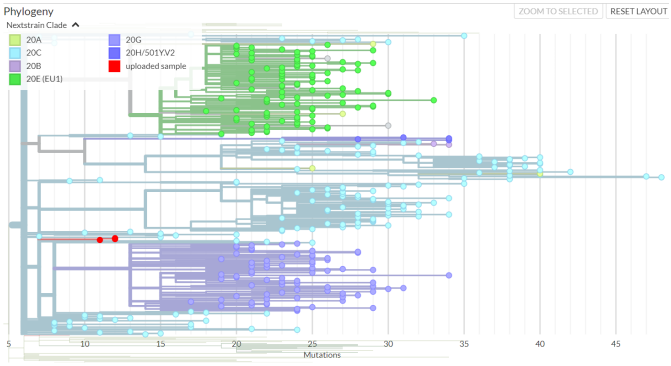

Fig. 4. Subtree from the SARS-CoV2 phylogeny with 2 samples added onto the 20C lineage[7].

7.3 UShER's Port as a Webtool UCSC Genome Browser

The UCSC Genome Browser is an interactive web tool created and hosted by the University of California, Santa Cruz. The Genome Browser was created in 2000 and had originally been developed for the purpose of storing and annotating the first draft of a fully assembled human genome sequence. The browser's use as a biological data visualization tool was thanks to its unprecedented ability to associate gene expression, conservation and regulation data with nucleotide precision. Since then, the UCSC Genome Browser has added the genomes of around 20 more different species with special genome assemblies for viruses such as SARS-CoV2 being added occasionally. The UCSC Genome Browser has also implemented join functionaly with numerous databases such as SwissProt for protein sequence data, OMIM (Online Mendelian Inheritance in Man) for phenotype data and SNPedia for SNP mutation data. With a reference genome for SARS-CoV2 already on the UCSC Genome Browser, creating a port of UShER on the Browser became a possibility. On the UCSC Genome browser UShER has found a use for placing sequences onto previously composed phylogenetic trees. However, unlike the Nextstrain port of UShER, the UCSC Genome Browser's port lacks the feature of transforming lineages found by UShER into joint temporal and spatial visualizations. Instead, the UCSC Genome Browser's port of

UShER outputs the results from computing a plausible lineage given a VCF file onto a custom session in the Genome Browser. This custom session is then populated with genome tracks corresponding to the positional data recorded on each sample featured in the VCF file given as input. In the UCSC Genome Browser, a track is defined as specific partition of some reference genome where abnormalities or unique features are expressed or likely to be expressed. With the creation of these tracks and custom sessions, a user would then be able to visualize the mutations characterizing the lineage computed by UShER. An example of the output produced by UShER's port on the UCSC Genome Browser is depicted in Figure 5. Figure 5 is a UCSC Genome Browser Session generated from the same VCF data used to generate the joint temporal and spatial visualization from Figure 4 on the Nextstrain port of UShER. In the session depicted by Figure 5, the coding regions of the reference SARS-CoV2 sequence s stored in the UCSC database is plotted in dark blue while the complete set of single nucleotide substitutions observed among the samples indicated in the input VCF file are shown as red and green vertical bars occupying certain indices on the plotted coding regions.

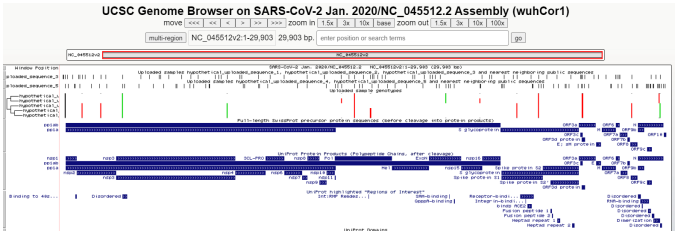

Fig. 5. Annotation on SARS-CoV2 reference genome given same VCF that was used as input in Figure 5. tree[10]

8 Conclusion

Considering the various aspects of phylogenetics involved in the study and handling of the 2019-2021 COVID-19 pandemic, the field of phylogenetics has clearly demonstrated its identity as an interdisciplinary field with a wide reach over epidemiology and computer science. The prolific involvement of phylogenetics in the development of studies and software for handling this pandemic has also demonstrated its applicability for maintaining everyday life. In addition to its large roles in the pandemic and contact tracing movement, phylogenetics remains a rapidly evolving field thanks to continuous advancements being made by institutes in academia as well as the advancements made in DNA sequencing technology. However, reflecting upon the development of UShER, it is important to recognize that even with such advancements being made so frequently and in large volume, great improvements in performance may only require a slight optimization or change in design choice to fulfill a specific role. This particular realization is coupled with the fact that close ties exist between phylogentics abstraction such as the phylogetic tree and lineage to computer science abstractions such as directed acyclic graphs. As the pandemic continues or new applications for phylogenetics are developed, good practice that researchers should adopt would be to ensure an initial understanding of the task from a biological point of view and then move onto actual feature engineering in the context of computer science. This was how the team behind UShER ended up succeeding and moving forward with this knowledge may set future project up for similarly successful results.

References

[1] Jialiang Yang et al. *Analysis on the reconstruction accuracy of the Fitch method for inferring ancestral states*, BMC Bioinformatics, 2013.[Online]. Available: *https* : *//www.ncbi.nlm.nih.gov/pmc/articles/PMC*3030536/

[2] José C Clemente et al. *Optimized ancestral state reconstruction using Sankoff parsimony*, BMC Bioinformatics, 2009.[Online]. Available: *https* : *//www.ncbi.nlm.nih.gov/pmc/articles/PMC*2677398/

[3] Kazutaka Katoh et al. *Adding unaligned sequences into an existing alignment using MAFFT and LAST*, Bioinformatics, 2012.[Online]. Available: *https* : *//pubmed.ncbi.nlm.nih.gov*/23023983/

[4] Bui Quang Minh et al. *IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era*, Mol Biol Evol., 2020.[Online]. Available: *https* : *//pubmed.ncbi.nlm.nih.gov*/32011700/

[5] Albert J Vilella et al. *EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates*, Genome Res., 2009.[Online]. Available: *https* : *//pubmed.ncbi.nlm.nih.gov*/19029536/

[6] Alexandros Stamatakis et al. *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies*, Bioinformatics, 2014.[Online]. Available: *https* : *//pubmed.ncbi.nlm.nih.gov*/24451623/

[7] J. Hadfield et al. *Nextstrain: real-time tracking of pathogen evolution*, Oxford Academic, 2018.[Online]. Available: $https://pubmed.ncbi.nlm.nih.gov/29790939/$

[8] R. Corbett-Detig et al. *Stability of SARS-CoV-2 phylogenies*, PLOS Genetics, 2020. [Online]. Available: $https://journals.plos.org/plosgenetics/article?id = 10.1371/journal.pgen.1009175$

[9] Ben Raphae. *Computational Methods for Biology Lecture 2* , Brown University 2009. [Online]. Available: $https://cs.brown.edu/courses/csci1950 - z/slides/CSCI1950ZFall09_Lecture2.pdf$

[10] Jim Kent et al. *UCSC Genome Browser on SARS-CoV-2 Jan. 2020* , UCSC 2021. [Online]. Available: $https://genome.ucsc.edu/$